WORLDROPE: PROBABILISTIC WORLD-CENTRIC ROTARY EMBEDDING FOR MULTI-VIEW ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Rotary Positional Embedding (RoPE) is widely used to encode relative positions in Transformers. In multi-view settings, however, patch coordinates are defined independently in each image: the same 3D point can therefore have different coordinates across views, making them insufficient for establishing cross-view correspondences. We introduce WorldRoPE, a probabilistic world-centric RoPE for multi-view attention. At each layer, WorldRoPE back-projects each patch token into a set of volumetric Gaussian primitives in a shared world frame. These primitives represent multiple plausible 3D supports for a token, jointly capturing alternative depth hypotheses, depth uncertainty, and the finite spatial extent of its patch frustum. When a key token's support is projected into a query view, it induces a distribution over where that key may appear in the query image. WorldRoPE integrates the rotary operator over this distribution in closed form, combining rotations at the projected primitive centers according to their mixture weights with covariance-dependent attenuation. This makes attention less sensitive to uncertain key locations in the query view, without rasterizing or warping token features. The formulation reduces to deterministic RoPE when the support collapses to a point. Across layers, the predicted geometry guides cross-view attention, while the updated features refine the geometry used at the next layer. Experiments on novel view synthesis, depth estimation, and 3D geometry reconstruction demonstrate the effectiveness of WorldRoPE for multi-view Transformers.

1 INTRODUCTION

Robotics and embodied perception require an understanding of the 3D world from observations taken at different viewpoints. Multi-view Transformers aggregate these observations through attention, with positional encoding introducing spatial information into token interactions. However, standard image-space Rotary Positional Embedding (RoPE) (Su et al., 2024; Heo et al., 2024) encodes relative displacement using coordinates defined independently within each image. The same 3D scene point can therefore have different coordinates across views, while nearby image locations may correspond to objects far apart in 3D. These image-space relationships alone are insufficient for establishing cross-view correspondences.

Recent camera-aware approaches incorporate viewing geometry into attention through relative camera poses and projective transformations, as in GTA (Miyato et al., 2024) and PRoPE (Li et al., 2025b). More recently, ray-based rotary encodings (Wu et al., 2026; Xie et al., 2026) further locate tokens along camera rays using predicted depths or predefined depth anchors and project these locations into the query view. An image patch, however, observes a volumetric region spanning its viewing frustum, rather than positions confined to camera rays, and may contain multiple objects at different depths. For example, a patch spanning a window frame and the scene behind it contains distinct depth cues, while the corresponding depth estimates may remain uncertain. In addition, the rays passing through a finite image patch span a volumetric frustum in 3D. A model of this 3D support should therefore jointly capture alternative depth hypotheses, depth uncertainty, and the finite spatial extent of the patch frustum.

To address these challenges, we introduce WorldRoPE, a probabilistic world-centric RoPE for multi-view Transformers. At each layer, we predict depth hypotheses, their uncertainty, and mixture weights from token features. Using these predictions, WorldRoPE back-projects each patch token

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of camera encodings for multi-view attention. Our method predicts depth hypotheses and their uncertainty for each token, representing possible 3D locations with volumetric Gaussian primitives. †: URoPE adopts predefined depth anchors across heads, while our method combines predicted hypotheses.

Method	Depth	Uncertainty	Multi-Hypothesis	Spatial Extent	Token Geometry
Plücker Ray (Jin et al., 2025)	–	–	–	–	Camera rays
GTA (Miyato et al., 2024)	–	–	–	–	–
PRoPE (Li et al., 2025b)	–	–	–	–	–
RayRoPE (Wu et al., 2026)	Predicted	✓	–	–	Ray segments
URoPE (Xie et al., 2026)	Fixed	–	✓†	–	Points on rays
WorldRoPE (ours)	Predicted	✓	✓	✓	**Volumetric Gaussians**

into a set of volumetric Gaussian primitives in the shared world space. These primitives represent multiple plausible 3D supports for a token, jointly capturing alternative depth hypotheses, depth uncertainty, and the finite spatial extent of its patch frustum. Projecting these primitives into the query view defines a distribution over where each key token may appear. We integrate the rotary operator over this distribution to guide cross-view attention, and the updated features inform geometry prediction at the next layer. Table 1 summarizes the differences from existing camera encodings.

We present a principled mathematical analysis of WorldRoPE and examine its properties in multi-view attention. First, the rotary expectation has a closed form that combines rotations at the projected primitive centers according to their mixture weights, with covariance-dependent attenuation. The resulting operator guides attention without explicitly rasterizing or warping token features. Second, the mixture operator is equivalent to a single scaled rotation for each rotary frequency block. The inverse of its effective attenuation can be interpreted as a token-adaptive temperature. Third, the formulation reduces to deterministic RoPE when the positional distribution collapses to a point. Finally, predicted geometry guides attention at each layer, while the updated features inform geometry prediction in the next layer.

Experiments on novel view synthesis, depth estimation, and 3D geometry reconstruction demonstrate the effectiveness of WorldRoPE for multi-view Transformers. Ablations and visualizations further examine the proposed geometric representation and rotary integration. Our main contributions are summarized as follows:

- We introduce WorldRoPE, a probabilistic world-centric RoPE that models token support with volumetric Gaussian primitives, jointly capturing alternative depth hypotheses, depth uncertainty, and the finite spatial extent of the patch frustum.
- We derive a closed-form rotary expectation over projected Gaussian distributions and integrate it into multi-view attention, coupling geometry prediction with feature updates across layers.
- Extensive experiments demonstrate the effectiveness of WorldRoPE across novel view synthesis, depth estimation, and 3D geometry reconstruction.

2 Related Work

Positional encodings for Transformers. Early approaches encode token positions using fixed or learned absolute embeddings (Vaswani et al., 2017; Dosovitskiy et al., 2021). Subsequent works (Liu et al., 2021; Press et al., 2022; Raffel et al., 2020; Shaw et al., 2018; Su et al., 2024) instead encode relative positions between tokens. Among them, RoPE (Su et al., 2024) encodes relative displacement by rotating queries and keys according to their coordinates, supporting generalization across positional ranges. Recent studies adapt RoPE to multidimensional visual coordinates for image recognition (Heo et al., 2024; Liu et al., 2026), 3D reconstruction (Zhang et al., 2024a; Ren et al., 2025), generation (Zhuo et al., 2024; Li et al., 2025a; Yang et al., 2025; Yesiltepe et al., 2026; Zhao et al., 2025), and video understanding (Wei et al., 2025). Other studies generalize RoPE using learned commuting matrices or Lie-group parameterizations (Yu et al., 2025; Ostmeier et al., 2025). While these methods apply positional transformations to deterministic token coordinates, WorldRoPE integrates RoPE over probabilistic token-position distributions for multi-view attention.

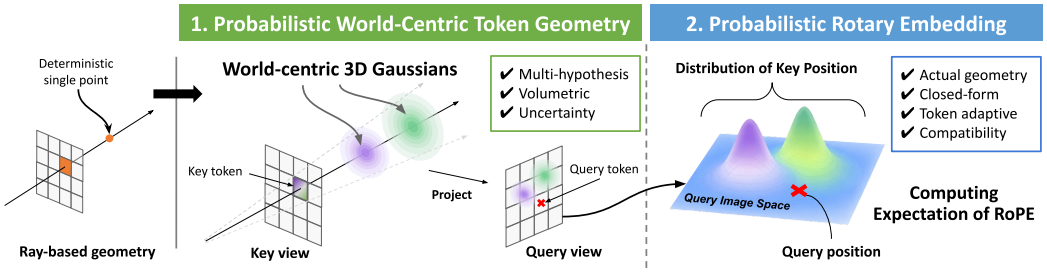


Figure 1: WorldRoPE overview. We model **probabilistic token positions** with **world-space Gaussian primitives**, capturing multiple depths, uncertainty, and patch extent. Their projections define possible key locations in the query view, over which **closed-form RoPE integration** guides multi-view attention without rasterizing or warping token features.

Camera-aware positional encodings for multi-view Transformers. Multi-view Transformers aggregate information across images for tasks such as novel view synthesis (Jin et al., 2025; Sajjadi et al., 2022; Venkat et al., 2023; Wang et al., 2025c; Kim et al., 2026), depth estimation, feed-forward 3D reconstruction (Jang et al., 2025; Keetha et al., 2026; Lin et al., 2025; Zhang et al., 2024b; Jia et al., 2025), and world modeling (Team HY-World, 2026; Team et al., 2026). To encode viewing geometry, earlier approaches concatenate camera raymaps or Plücker raymaps with image features (Jin et al., 2025). Recent approaches integrate camera information into RoPE to encode spatial relationships across views. GTA (Miyato et al., 2024) and PRoPE (Li et al., 2025b) directly transform token features using camera extrinsic matrices and projection matrices, respectively. More recently, URoPE (Xie et al., 2026) and RayRoPE (Wu et al., 2026) use ray-based geometry for rotary encoding in the query camera frame, with predefined anchors and predicted depth, respectively. In contrast, WorldRoPE represents the 3D support of each token as a mixture of volumetric Gaussian primitives in a shared world frame.

3 WorldRoPE

We propose WorldRoPE, a probabilistic world-centric Rotary Positional Embedding (RoPE) for multi-view attention, as shown in Figure 1. After reviewing notation and RoPE in Section 3.1, we model probabilistic token geometry with world-space Gaussian primitives in Section 3.2. We then project these primitives into the query view and integrate the rotary operator over their positional distributions in Section 3.3. Finally, we analyze the properties of WorldRoPE in Section 3.4.

3.1 Preliminaries

Notation. Consider a patch j in a query view q and a patch i in a key view k. Their coordinates are $\mathbf{u}_{q,j} = (u_{q,j}, v_{q,j})^\top$ and $\mathbf{u}_{k,i} = (u_{k,i}, v_{k,i})^\top$. For each view r, let $\mathbf{K}_r$ and $(\mathbf{R}_r, \mathbf{t}_r)$ denote its intrinsic and extrinsic camera parameters. At layer ℓ, $\mathbf{f}_{r,s}^{(\ell)}$ denotes the feature of patch s in view r before the attention projections. Suppressing the layer index for brevity, we compute $\mathbf{q}_{q,j} = \mathbf{W}_Q \mathbf{f}_{q,j} \in \mathbb{R}^d$, $\mathbf{k}_{k,i} = \mathbf{W}_K \mathbf{f}_{k,i} \in \mathbb{R}^d$, and $\mathbf{v}_{k,i} = \mathbf{W}_V \mathbf{f}_{k,i} \in \mathbb{R}^d$, where d is the attention-head dimension.

Rotary Positional Embedding (RoPE). Standard RoPE (Su et al., 2024; Heo et al., 2024) encodes relative positions by rotating queries and keys before computing their inner product. For a 2D image coordinate $\mathbf{u} = (u, v)^\top$, the bth channel pair is rotated by $\boldsymbol{\omega}_b^\top \mathbf{u}$, where $\boldsymbol{\omega}_b \in \mathbb{R}^2$ is its frequency vector. The full operator is given by

$$\mathcal{R}(\mathbf{u}) = \mathrm{blockdiag}\Big(R(\boldsymbol{\omega}_1^\top \mathbf{u}), \ldots, R(\boldsymbol{\omega}_{d/2}^\top \mathbf{u})\Big), \quad \text{where} \quad R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{1}$$

Applying this operator gives the rotated query and key, denoted by $\mathcal{R}(\mathbf{u}_{q,j})\mathbf{q}_{q,j}$ and $\mathcal{R}(\mathbf{u}_{k,i})\mathbf{k}_{k,i}$, respectively. When the coordinates are expressed in a common image frame, the inner product of the rotated query and key depends only on their relative displacement, since $\mathcal{R}(\mathbf{u}_{q,j})^\top \mathcal{R}(\mathbf{u}_{k,i}) = \mathcal{R}(\mathbf{u}_{k,i} - \mathbf{u}_{q,j})$. In the following, we omit the frequency index b when discussing a single block.

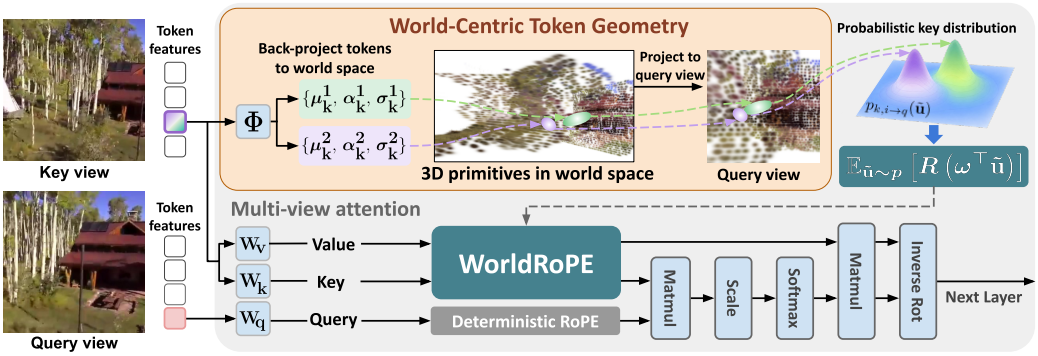


Figure 2: Multi-view attention with WorldRoPE. At each attention layer, we predict world-space Gaussian primitives from token features. Projecting them into the query view places key and query positions in a common image frame. We integrate RoPE over the projected key distribution to guide cross-view attention, and the updated features guide geometry prediction in the next layer.

3.2 Probabilistic World-Space Token Geometry

Multi-hypothesis depth prediction. At each layer, a prediction head Φ estimates a depth distribution along the camera ray passing through the patch center. We model this distribution as an M-component Gaussian mixture with parameters given by

$$\left\{(\alpha_{k,i}^m, \mu_{k,i}^m, \sigma_{k,i}^m)\right\}_{m=1}^M = \Phi(\mathbf{f}_{k,i}), \tag{2}$$

where m indexes depth hypotheses, and $\mu_{k,i}^m$ and $\sigma_{k,i}^m$ denote the mean and standard deviation of the depth along the patch-center ray. The weights satisfy $\alpha_{k,i}^m \geq 0$ and $\sum_m \alpha_{k,i}^m = 1$. This mixture retains alternative depths and their uncertainty for patches containing objects at different depths.

Volumetric Gaussian primitives. An often overlooked fact is that a 2D image token corresponds to a volumetric region within the patch viewing frustum, extending beyond its center ray. We therefore construct a Gaussian primitive around each depth hypothesis to capture both the patch extent and depth uncertainty. Assuming independent pixel and depth offsets, we linearize back-projection to obtain the Gaussian mean and approximate covariance in the key-camera frame as follows:

$$\boldsymbol{\mu}_{k,i}^m = \mu_{k,i}^m \mathbf{r}_{k,i} \quad \text{and} \quad \boldsymbol{\Sigma}_{k,i}^m \approx (\mu_{k,i}^m)^2 \mathbf{B}_k \mathbf{C} \mathbf{B}_k^\top + (\sigma_{k,i}^m)^2 \mathbf{r}_{k,i} \mathbf{r}_{k,i}^\top, \tag{3}$$

where $\mathbf{r}_{k,i} = \mathbf{K}_k^{-1}(u_{k,i}, v_{k,i}, 1)^\top$ and $\mathbf{B}_k$ contains the first two columns of $\mathbf{K}_k^{-1}$. For uniform pixel offsets within a patch of size $\Delta_u \times \Delta_v$, $\mathbf{C} = \mathrm{diag}(\Delta_u^2/12, \Delta_v^2/12)$. The two covariance terms capture the patch extent and depth uncertainty, respectively. See Appendix B.3 for the derivation.

World-space Gaussian support. We transform the primitives into a shared 3D world space using the key-camera extrinsics $(\mathbf{R}_k, \mathbf{t}_k)$ as follows:

$$\boldsymbol{\mu}_{k,i}^{\mathrm{w},m} = \mathbf{R}_k^\top (\boldsymbol{\mu}_{k,i}^m - \mathbf{t}_k) \quad \text{and} \quad \boldsymbol{\Sigma}_{k,i}^{\mathrm{w},m} = \mathbf{R}_k^\top \boldsymbol{\Sigma}_{k,i}^m \mathbf{R}_k, \tag{4}$$

where $\boldsymbol{\mu}_{k,i}^{\mathrm{w},m}$ and $\boldsymbol{\Sigma}_{k,i}^{\mathrm{w},m}$ are the world-space mean and covariance. The primitives share the token feature $\mathbf{f}_{k,i}$, from which the key and value vectors are computed, and retain their mixture weights $\alpha_{k,i}^m$. Together, they represent multiple plausible 3D supports for the token, with positional density given by

$$p_{k,i}^{\mathrm{w}}(\mathbf{x}) = \sum_{m=1}^M \alpha_{k,i}^m \mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}_{k,i}^{\mathrm{w},m}, \boldsymbol{\Sigma}_{k,i}^{\mathrm{w},m}\right), \tag{5}$$

where $\mathbf{x} \in \mathbb{R}^3$ denotes a world-space position. This mixture probabilistically represents where a 2D image token may lie in 3D world space using a set of volumetric Gaussian primitives.

3.3 PROBABILISTIC ROTARY POSITIONAL EMBEDDING

Projection into the query view. We project the world-space Gaussian primitives into the query view to determine where each key token may appear. Following 3D Gaussian Splatting (Kerbl et al., 2023), a local affine approximation yields the projected mean and covariance as follows:

$$\boldsymbol{\mu}^m_{k,i\to q} = \pi_q\Big(\mathbf{R}_q\boldsymbol{\mu}^{\mathrm{w},m}_{k,i} + \mathbf{t}_q\Big), \qquad \boldsymbol{\Sigma}^m_{k,i\to q} = \mathbf{J}^m_{k,i\to q}\mathbf{R}_q\boldsymbol{\Sigma}^{\mathrm{w},m}_{k,i}\mathbf{R}_q^\top(\mathbf{J}^m_{k,i\to q})^\top, \tag{6}$$

where π_q denotes perspective projection from query-camera coordinates to pixel coordinates using $\mathbf{K}_q$. Its Jacobian $\mathbf{J}^m_{k,i\to q} \in \mathbb{R}^{2\times 3}$ is computed at the primitive center in the query-camera frame. The projected primitives form a Gaussian mixture with positional density given by

$$p_{k,i\to q}(\tilde{\mathbf{u}}) = \sum_{m=1}^{M} \alpha^m_{k,i}\mathcal{N}\big(\tilde{\mathbf{u}}; \boldsymbol{\mu}^m_{k,i\to q}, \boldsymbol{\Sigma}^m_{k,i\to q}\big), \tag{7}$$

where $\tilde{\mathbf{u}} \in \mathbb{R}^2$ denotes a possible projected key position in the query image. This distribution captures the possible key locations and their uncertainty without rasterizing or warping token features.

Closed-form rotary integration. Unlike the deterministic positions used in standard RoPE, our projected key positions are probabilistic. We therefore integrate the rotary operator over the projected distribution, defining $\overline{\mathcal{R}}_{k,i\to q} := \mathbb{E}_{\tilde{\mathbf{u}}\sim p_{k,i\to q}}[\mathcal{R}(\tilde{\mathbf{u}})]$. We compute this expectation for each 2×2 rotation block, $R(\boldsymbol{\omega}^\top\tilde{\mathbf{u}})$; applying the result independently to all blocks yields the full operator $\mathcal{R}$. For a fixed frequency $\boldsymbol{\omega}$, the Gaussian expectation is given by

$$\begin{aligned}\mathbb{E}_{\tilde{\mathbf{u}}\sim\mathcal{N}(\boldsymbol{\mu}^m_{k,i\to q},\boldsymbol{\Sigma}^m_{k,i\to q})}[R(\boldsymbol{\omega}^\top\tilde{\mathbf{u}})] &= \int_{\mathbb{R}^2} R(\boldsymbol{\omega}^\top\tilde{\mathbf{u}})\mathcal{N}\big(\tilde{\mathbf{u}}; \boldsymbol{\mu}^m_{k,i\to q}, \boldsymbol{\Sigma}^m_{k,i\to q}\big)\, d\tilde{\mathbf{u}} \\ &= \underbrace{\exp\big(-\tfrac{1}{2}\boldsymbol{\omega}^\top\boldsymbol{\Sigma}^m_{k,i\to q}\boldsymbol{\omega}\big)}_{a^m_{k,i\to q}} \quad \underbrace{R(\boldsymbol{\omega}^\top\boldsymbol{\mu}^m_{k,i\to q})}_{\text{rotation at the projected mean}}.\end{aligned} \tag{8}$$

Each primitive contributes a rotation at its projected center with covariance-dependent attenuation $a^m_{k,i\to q}$. By linearity of expectation, the mixture combines these contributions into an expected block, which can be written as a single scaled rotation as follows:

$$\overline{R}_{k,i\to q} = \sum_{m=1}^{M} \alpha^m_{k,i}a^m_{k,i\to q}R(\boldsymbol{\omega}^\top\boldsymbol{\mu}^m_{k,i\to q}) = \bar{a}_{k,i\to q}R(\bar{\theta}_{k,i\to q}), \tag{9}$$

where $\bar{a}_{k,i\to q}$ and $\bar{\theta}_{k,i\to q}$ denote the effective attenuation and phase for the current frequency block. Placing these blocks along the diagonal yields the full operator $\overline{\mathcal{R}}_{k,i\to q} \in \mathbb{R}^{d\times d}$. This expectation is exact under the projected Gaussian approximation. See Appendix B.1 for the derivation.

Multi-view attention. We apply RoPE to the query at its deterministic coordinate $\mathbf{u}_{q,j}$ and the expected rotary operator to each key in query view q, as shown in Figure 2. The attention score and weight are given by

$$s^{q\leftarrow k}_{ji} = \frac{\big\langle\mathcal{R}(\mathbf{u}_{q,j})\mathbf{q}_{q,j}, \overline{\mathcal{R}}_{k,i\to q}\mathbf{k}_{k,i}\big\rangle}{\sqrt{d}}, \qquad A^{q\leftarrow k}_{ji} = \mathrm{softmax}_{(k,i)}\Big(s^{q\leftarrow k}_{ji}\Big), \tag{10}$$

where softmax normalizes over all attended key tokens (k, i) across views. Following GTA (Miyato et al., 2024), we also apply positional transformations to value and output features, as detailed in Appendix A.7. The features updated via attention then guide geometry prediction at the next layer.

3.4 PROPERTIES OF WORLDROPE

Token-adaptive temperature. The inverse of the effective attenuation can be interpreted as a token-adaptive temperature for each rotary frequency block. Let $\mathbf{q}, \mathbf{k} \in \mathbb{R}^2$ denote the query and key channel pairs at frequency $\boldsymbol{\omega}$. For $\bar{a}_{k,i\to q} > 0$, their score contribution can be rewritten:

$$\frac{\big\langle R(\boldsymbol{\omega}^\top\mathbf{u}_{q,j})\mathbf{q}, \overline{R}_{k,i\to q}\mathbf{k}\big\rangle}{\sqrt{d}} = \frac{\big\langle R(\boldsymbol{\omega}^\top\mathbf{u}_{q,j})\mathbf{q}, R(\bar{\theta}_{k,i\to q})\mathbf{k}\big\rangle}{\sqrt{d}\,\tau_{k,i\to q}}, \tag{11}$$

where $\tau_{k,i\to q} = 1/\bar{a}_{k,i\to q}$ acts as a token-adaptive temperature. The effective attenuation depends on the covariance of each hypothesis; a smaller $\bar{a}_{k,i\to q}$ corresponds to a higher temperature, reducing the magnitude of the score contribution for that frequency block.

Table 2: Comparison of positional encodings for novel view synthesis with LVSM on RE10K and Objaverse. WorldRoPE achieves the best performance across all metrics on both datasets.

Method	**RE10K** (Zhou et al., 2018)			**Objaverse** (Deitke et al., 2023)		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Plücker Ray (Jin et al., 2025)	24.62	0.790	0.108	15.15	0.838	0.355
GTA (Miyato et al., 2024)	25.05	0.799	0.102	22.22	0.892	0.123
PRoPE (Li et al., 2025b)	25.44	0.811	0.095	22.41	0.894	0.117
RayRoPE (Wu et al., 2026)	26.17	0.833	0.081	22.25	0.892	0.121
URoPE (Xie et al., 2026)	26.26	0.834	0.078	21.99	0.890	0.128
WorldRoPE (ours)	**26.75**	**0.846**	**0.073**	**22.52**	**0.896**	**0.113**

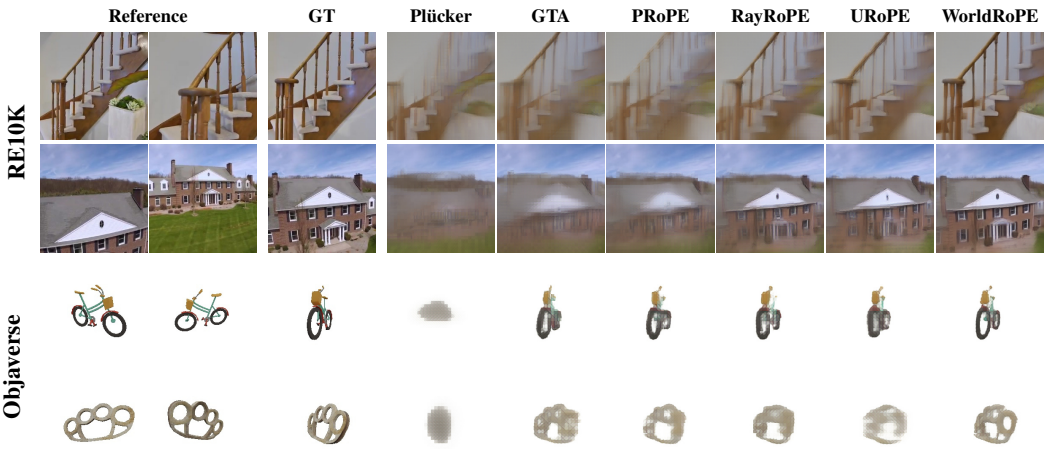


Figure 3: Qualitative comparison of novel view synthesis on RE10K and Objaverse. WorldRoPE better preserves thin structures and fine details like stair railings and bicycle wheels.

Sensitivity to positional displacement. Our covariance-dependent attenuation reduces sensitivity to errors in the projected centers of primitives. Consider an arbitrary projected Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with attenuation $a = \exp(-\boldsymbol{\omega}^\top \boldsymbol{\Sigma}\boldsymbol{\omega}/2)$ at frequency $\boldsymbol{\omega}$. The operator change under a center displacement $\boldsymbol{\delta} \in \mathbb{R}^2$ is bounded as follows:

$$\left\| aR\big(\boldsymbol{\omega}^\top(\boldsymbol{\mu}+\boldsymbol{\delta})\big) - aR(\boldsymbol{\omega}^\top\boldsymbol{\mu}) \right\|_2 \le a\left|\boldsymbol{\omega}^\top\boldsymbol{\delta}\right|, \tag{12}$$

where $\|\cdot\|_2$ denotes the spectral norm. Greater positional uncertainty, measured by $\boldsymbol{\omega}^\top\boldsymbol{\Sigma}\boldsymbol{\omega}$, decreases the attenuation factor a. This lowers the upper bound on sensitivity to projected-center errors. See Appendix B.1 for the derivation.

Deterministic RoPE as a limiting case. Our formulation recovers deterministic RoPE when one hypothesis $m^\star$ has unit weight and its patch extent and depth uncertainty vanish as follows:

$$\alpha_{k,i}^{m^\star} \to 1, \quad \boldsymbol{\Sigma}_{k,i\to q}^{m^\star} \to \mathbf{0} \quad \Longrightarrow \quad \overline{\mathcal{R}}_{k,i\to q} \to \mathcal{R}(\boldsymbol{\mu}_{k,i\to q}^{m^\star}). \tag{13}$$

Thus, WorldRoPE extends rotary encoding to probabilistic positional distributions while retaining deterministic RoPE as a limiting case.

4 Experiments

We first evaluate WorldRoPE for novel view synthesis against existing positional encodings. We then conduct detailed analyses of WorldRoPE, including ablation studies. Finally, we demonstrate that WorldRoPE can serve as a plug-and-play mechanism for other tasks using multi-view attention.

4.1 Main Results

Experimental setup. We test WorldRoPE within LVSM (Jin et al., 2025) on RE10K (Zhou et al., 2018) and Objaverse (Deitke et al., 2023). We compare with Plücker ray encoding (Jin et al., 2025), GTA (Miyato et al., 2024), PRoPE (Li et al., 2025b), RayRoPE (Wu et al., 2026), and URoPE (Xie et al., 2026), using PSNR, SSIM, and LPIPS (Zhang et al., 2018) to measure rendered image quality.

Table 3: Component ablation on RE10K. Removing any component degrades performance; modeling with probabilistic primitives yields the largest performance gain.

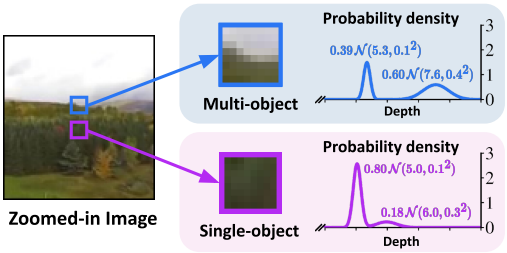
Method	PSNR ↑	SSIM ↑	LPIPS ↓
w/o Probabilistic modeling	26.30	0.835	0.082
w/o Multi-hypothesis	26.50	0.839	0.077
w/o Patch footprint	26.68	0.845	0.074
WorldRoPE	**26.75**	**0.846**	**0.073**

Table 4: Effect of the number of hypotheses on RE10K. Performance improves from one to four hypotheses, while eight brings no further gain. We use four hypotheses per token.

# of hypotheses	PSNR ↑	SSIM ↑	LPIPS ↓
1	26.50	0.839	0.077
2	26.61	0.842	0.075
4	**26.75**	**0.846**	**0.073**
8	26.73	**0.846**	**0.073**



(a) Tokens with multiple depth peaks



(b) Predicted depth distributions for each token

Figure 4: Visualization of predicted depth hypotheses. (a) Red highlights tokens with multiple prominent peaks in their depth distributions. (b) A patch spanning vegetation and sky exhibits two distinct peaks, while a patch within vegetation has one dominant peak.

Quantitative results. Table 2 shows that WorldRoPE consistently outperforms all baselines across all metrics on both the RE10K dataset of real-world scenes and the Objaverse dataset of object-centric scenes. The improvement is particularly pronounced on the RE10K dataset, with a PSNR gain of 0.49 dB over the strongest baseline. This suggests that modeling alternative depths and their spatial extent is particularly useful for scenes containing foreground and background structures, where a patch can span multiple depths and correspond to a broader region in another view.

Qualitative results. Figure 3 compares synthesized target views. WorldRoPE better preserves thin structures and fine details, including stair railings, window boundaries, bicycle wheels, and brass knuckles, which the baselines often blur or merge with their surroundings. These observations suggest that modeling alternative depth hypotheses helps distinguish structures at different depths within a patch. The qualitative improvements are consistent with the gains in PSNR, SSIM, and LPIPS.

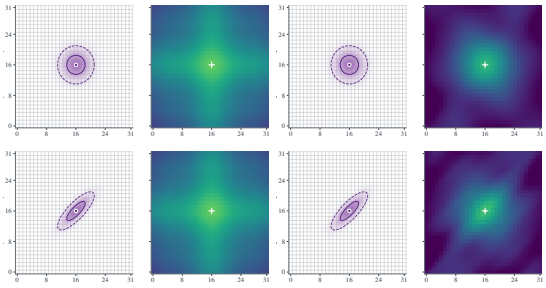
4.2 Analysis and Discussion

Component ablation. Table 3 separates the effects of probabilistic modeling, multiple depth hypotheses, and the patch footprint. Removing probabilistic modeling causes the largest drop, indicating the importance of incorporating positional uncertainty into attention. Performance also declines with a single hypothesis, supporting the representation of alternative depths within a patch. The smaller but consistent drop without the patch footprint shows that its spatial extent contributes beyond the patch-center ray.

Number of hypotheses. Table 4 tests how many depth hypotheses are needed per token. Increasing the number from one to four improves all three metrics, supporting the use of alternative depths within a patch. Eight hypotheses bring no further gain; we therefore use four primitives per token.

Visualization of depth hypotheses. Figure 4 examines how predicted depth hypotheses vary with patch content. In Figure 4a, red highlights tokens with multiple peaks, which often appear near object boundaries and overlapping structures. A patch spanning vegetation and sky produces two peaks, whereas a patch within vegetation produces one dominant peak, as shown in Figure 4b. This contrast shows that WorldRoPE can preserve distinct depths within a patch instead of collapsing them into a single estimate.

Compatibility with other rotary operators. To examine how the rotary operator uses our predicted Gaussian covariance, we replace the axial operator used in the main experiments with a mixed

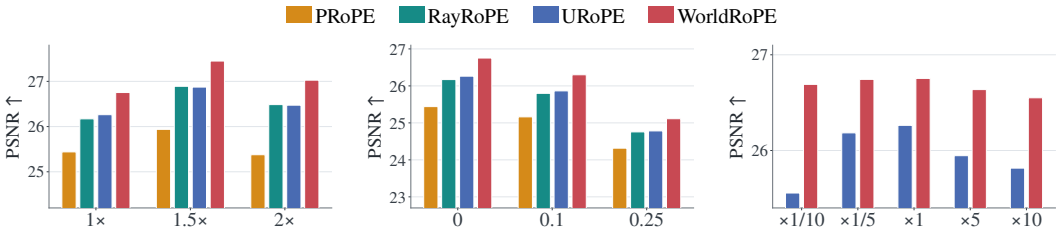


(a) Axial rotary (b) Mixed rotary

Figure 5: Projected Gaussians and their responses under axial and mixed rotary operators. Only the mixed operator responds to off-diagonal covariance.

Table 5: PSNR with axial and mixed rotary operators on Objaverse. Parenthesized values denote changes from the axial variant. The mixed operator yields the largest gain for WorldRoPE. Additional results are reported in Appendix A.1.

Method	Operator	PSNR ↑
RayRoPE (Wu et al., 2026)	Axial	22.25
	Mixed	22.33 (+0.08)
URoPE (Xie et al., 2026)	Axial	21.99
	Mixed	22.03 (+0.04)
WorldRoPE (ours)	Axial	**22.52**
	Mixed	**22.71** (+0.19)



(a) Zoom-in (b) Rotation noise (c) Depth initialization

Figure 6: Robustness to (a) camera zoom and (b) rotation noise during evaluation, and (c) depth configuration during training. WorldRoPE achieves robustness across all tested settings.

operator. Axial operators use frequency vectors aligned with one image axis, whereas mixed operators combine both coordinates.[1] In the Gaussian expectation derived in Section 3.3, this distinction allows only mixed operators to use covariance between the image coordinates, as illustrated in Figures 5a and 5b. On the Objaverse dataset, the mixed operator improves WorldRoPE by 0.19 dB in PSNR, compared with 0.08 dB for RayRoPE and 0.04 dB for URoPE (Table 5). The larger gain suggests that incorporating off-diagonal covariance into the rotary response makes our Gaussian modeling more effective. See Appendix A.1 for results on the RE10K dataset.

Robustness. We test sensitivity to camera perturbations during evaluation and depth initialization during training. Under camera zoom and rotation noise, WorldRoPE achieves higher PSNR than the baselines (Figures 6a and 6b). Zoom changes the projected extent of a patch, whereas rotation noise shifts its projected position. These results are consistent with the benefits of modeling the patch footprint and accounting for positional uncertainty. We also scale the predefined depth anchors of URoPE and the initial depth biases of WorldRoPE from 0.1 to 10 times their default values. WorldRoPE maintains higher PSNR with less variation across scales (Figure 6c). Its depth hypotheses are updated from token features during training, reducing dependence on their initial placement.

Layerwise geometry refinement. Figure 7 shows the predicted Gaussian primitives across attention layers and their projection into the query view. In deeper layers, their smaller extents indicate reduced positional uncertainty, while the scene structure becomes more recognizable. The projected key primitives approximately align with the query image, although their centers no longer follow a regular image grid and their regions vary in size and orientation. Some regions are broader near image boundaries, consistent with perspective magnification of a key-view patch when projected into the query view. See Appendix A.4 for additional visualizations across all attention layers.

4.3 Broader Applications

We further show that WorldRoPE can serve as a plug-and-play mechanism for camera-conditioned multi-view attention, including depth estimation and feed-forward 3D reconstruction.

[1] Axial operators use $\boldsymbol{\omega} = (\omega_u, 0)^\top$ or $(0, \omega_v)^\top$, whereas mixed operators use $\boldsymbol{\omega} = (\omega_u, \omega_v)^\top$.

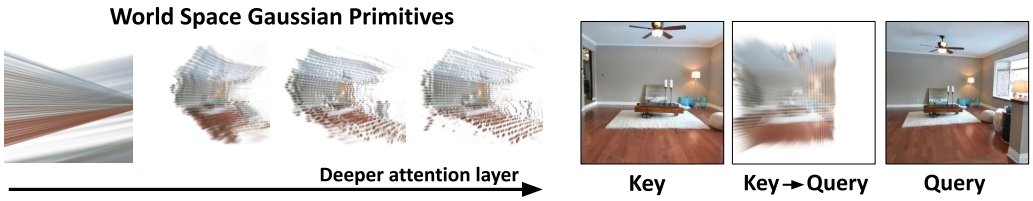


Figure 7: World-space Gaussian primitives across attention layers (left) and their projection from the key view into the query view (right). RGB rasterization is used only for visualization; the primitives carry token features and are not rasterized during attention.

Table 6: Comparison of camera-aware positional encodings for stereo depth estimation with UniMatch. We report depth errors using AbsRel and RMSE on the Scenes11, SUN3D, and RGBD datasets, and assess cross-dataset generalization on the unseen ScanNet++ dataset. WorldRoPE outperforms existing positional encodings in RMSE on all four datasets.

Method	Scenes11		SUN3D		RGBD		ScanNet++ (unseen)	
	Abs Rel ↓	RMSE ↓	Abs Rel ↓	RMSE ↓	Abs Rel ↓	RMSE ↓	Abs Rel ↓	RMSE ↓
w/o camera encoding	0.176	1.182	0.179	0.524	0.185	0.953	0.286	0.482
Plücker Ray (Jin et al., 2025)	0.123	0.943	0.130	0.391	0.145	0.790	0.276	0.464
GTA (Miyato et al., 2024)	0.116	0.830	0.133	0.384	0.134	0.693	0.325	0.520
PRoPE (Li et al., 2025b)	0.106	0.778	0.136	0.390	0.134	0.695	0.263	0.434
RayRoPE (Wu et al., 2026)	0.101	0.743	0.125	0.370	0.128	0.677	0.257	0.426
URoPE (Xie et al., 2026)	0.093	0.686	0.123	0.374	**0.119**	0.637	0.234	0.398
WorldRoPE (ours)	**0.079**	**0.661**	**0.116**	**0.356**	**0.119**	**0.630**	**0.224**	**0.386**

Depth estimation. We integrate camera-aware positional encodings into the cross-attention layers of UniMatch (Xu et al., 2023) and compare their depth estimation performance. We train them on the training splits of the Scenes11 (Ummenhofer et al., 2017), SUN3D (Xiao et al., 2013), and RGBD (Sturm et al., 2012) datasets, and evaluate on their test splits, following UniMatch. We further evaluate cross-dataset generalization on the unseen ScanNet++ dataset (Yeshwanth et al., 2023). Table 6 shows that WorldRoPE achieves the lowest absolute relative error (Abs Rel) and root mean squared error (RMSE) for depth estimation across all four datasets. The improvements in both metrics on the unseen dataset further support generalization to unseen scenes. Further implementation details and qualitative results are provided in Sections C.3 and D.2, respectively.

Feed-forward 3D reconstruction. We fine-tune pretrained VGGT (Wang et al., 2025a) by adding attention layers with camera-aware positional encodings. We train on the ScanNet++ (Yeshwanth et al., 2023) training split and evaluate on its test split. Table 7 reports depth errors (AbsRel and normalized RMSE), camera-space point AbsRel, and camera rotation error. WorldRoPE outperforms baselines on all four metrics, demonstrating that our world-centric positional encoding can be effectively applied when fine-tuning a pretrained model. Further implementation details are shown in Appendix C.4.

Table 7: Comparison of camera-aware positional encodings for feed-forward 3D reconstruction with VGGT. WorldRoPE achieves the lowest depth, camera-space point, and camera rotation errors.

Method	ScanNet++			
	Depth	Norm.	Cam.-pts	Rot.
	AbsRel ↓	RMSE ↓	AbsRel ↓	Err. ↓
RayRoPE	0.0579	0.0624	0.0575	1.04
URoPE	0.0581	0.0626	0.0576	1.02
WorldRoPE	**0.0567**	**0.0619**	**0.0565**	**0.98**

5 Conclusion

We presented WorldRoPE, a probabilistic world-centric RoPE for multi-view attention. Our method represents token geometry with volumetric Gaussian primitives that jointly capture alternative depth hypotheses, depth uncertainty, and the spatial extent of the patch frustum. Integrating the rotary operator over their projected distributions yields a closed-form encoding that attenuates uncertain positional interactions without rasterizing or warping token features. Across layers, predicted geometry guides attention, while updated features inform subsequent geometry predictions. Experiments on novel view synthesis, depth estimation, and feed-forward 3D reconstruction demonstrate the effectiveness of WorldRoPE across multi-view tasks and architectures.

AI USE STATEMENT

We used generative AI tools as part of our workflow. These tools helped us discuss methodological choices and mathematical claims, interpret experimental results, identify relevant literature, and draft and refine portions of the manuscript. They were also used to assist with graph preparation and figure layout. All reported results and graphs are based on experiments designed, conducted, and verified by the authors. The authors reviewed and revised all AI-assisted material and take full responsibility for the final text, technical claims, references, figures, and experimental results.

ETHICS STATEMENT

This work evaluates the proposed method on existing datasets and does not introduce new data collection. Like other multi-view reconstruction methods, it can recover scene structure and visual details from images observed across multiple viewpoints. When applied to private or sensitive imagery, its use should respect the consent of relevant individuals, dataset licenses and usage restrictions, and appropriate privacy safeguards. We do not anticipate additional ethical concerns beyond those associated with the source datasets and potential applications of multi-view reconstruction.

REPRODUCIBILITY STATEMENT

The main paper describes the proposed world-space token representation and its closed-form rotary integration. The appendix provides further derivations, including the construction and projection of the Gaussian primitives, as well as analyses of the resulting rotary operators. For each task, the experimental sections and appendix specify the model adaptation, training configuration, dataset splits, evaluation protocol, and relevant implementation details. They also report ablations that isolate the effects of multi-hypothesis modeling, patch footprint, uncertainty-dependent attenuation, and the number of primitives.

REFERENCES

Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, 2021.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *ECCV*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Wonbong Jang, Philippe Weinzaepfel, Vincent Leroy, Lourdes Agapito, and Jerome Revaud. Pow3r: Empowering unconstrained 3d reconstruction with camera and scene priors. In *CVPR*, 2025.

Heng Jia, Linchao Zhu, and Na Zhao. H3r: hybrid multi-view correspondence for generalizable 3d reconstruction. In *ICCV*, 2025.

Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. In *ICLR*, 2025.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *3DV*. IEEE, 2026.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 2023.

Evan Kim, Hyunwoo Ryu, Thomas W Mitchel, and Vincent Sitzmann. Scaling view synthesis transformers. In *CVPR*, 2026.

Jiaye Li, Baoyou Chen, Hui Li, Zilong Dong, Jingdong Wang, and Siyu Zhu. Head-wise adaptive rotary positional encoding for fine-grained image generation. *arXiv preprint arXiv:2510.10489*, 2025a.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. In *NeurIPS*, 2025b.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Haoyu Liu, Sucheng Ren, Tingyu Zhu, Peng Wang, Cihang Xie, Alan Yuille, Zeyu Zheng, and Feng Wang. Spiral rope: Rotate your rotary positional embeddings in the 2d plane. *arXiv preprint arXiv:2602.03227*, 2026.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

Takeru Miyato, Bernhard Jaeger, Max Welling, and Andreas Geiger. Gta: A geometry-aware attention mechanism for multi-view transformers. In *ICLR*, 2024.

Sophie Ostmeier, Brian Axelrod, Maya Varma, Michael E Moseley, Akshay Chaudhari, and Curtis Langlotz. Liere: Lie rotational positional encodings. In *ICML*, 2025.

Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *ICLR*, 2022.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020.

Jiahui Ren, Mochu Xiang, Jiajun Zhu, and Yuchao Dai. Panosplatt3r: Leveraging perspective pretraining for generalized unposed wide-baseline panorama reconstruction. In *ICCV*, 2025.

Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *CVPR*, 2022.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *NAACL-HLT*, 2018.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *IROS*, 2012.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

Team HY-World. Hy-world 2.0: A multi-modal world model for reconstructing, generating, and simulating 3d worlds. *arXiv preprint arXiv:2604.14268*, 2026. doi: 10.48550/arXiv.2604.14268. URL https://arxiv.org/abs/2604.14268.

Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. DeMoN: Depth and motion network for learning monocular stereo. In *CVPR*, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

Naveen Venkat, Mayank Agarwal, Maneesh Singh, and Shubham Tulsiani. Geometry-biased transformers for novel view synthesis. *arXiv preprint arXiv:2301.04650*, 2023.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025a.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025b.

Zhengqing Wang, Yuefan Wu, Jiacheng Chen, Fuyang Zhang, and Yasutaka Furukawa. Clift: Compressive light-field tokens for compute efficient and adaptive neural rendering. In *NeurIPS*, 2025c.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, et al. Videorope: What makes for good video rotary position embedding? In *ICML*, 2025.

Yu Wu, Minsik Jeon, Jen-Hao Rick Chang, Oncel Tuzel, and Shubham Tulsiani. Rayrope: Projective ray positional encoding for multi-view attention. In *ECCV*, 2026.

Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *ICCV*, 2013.

Yichen Xie, Depu Meng, Chensheng Peng, Yihan Hu, Quentin Herau, Masayoshi Tomizuka, and Wei Zhan. Urope: Universal relative position embedding across geometric spaces. In *ECCV*, 2026.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

Hidir Yesiltepe, Tuna Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. In *CVPR*, 2026.

Hao Yu, Tangyu Jiang, Shuning Jia, Shannan Yan, Shunning Liu, Haolong Qian, Guanghao Li, Shuting Dong, and Chun Yuan. Comrope: Scalable and robust rotary position embedding parameterized by trainable commuting angle matrices. In *CVPR*, 2025.

Chubin Zhang, Hongliang Song, Yi Wei, Yu Chen, Jiwen Lu, and Yansong Tang. Geolrm: Geometry-aware large reconstruction model for high-quality 3d gaussian generation. In *NeurIPS*, 2024a.

Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *ECCV*, 2024b.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Min Zhao, Guande He, Yixiao Chen, Hongzhou Zhu, Chongxuan Li, and Jun Zhu. Riflex: A free lunch for length extrapolation in video diffusion transformers. In *ICML*, 2025.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics (TOG)*, 2018.

Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Xiangyang Zhu, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. In *NeurIPS*, 2024.

A Additional Analysis

A.1 Compatibility with Mixed Rotary Operator

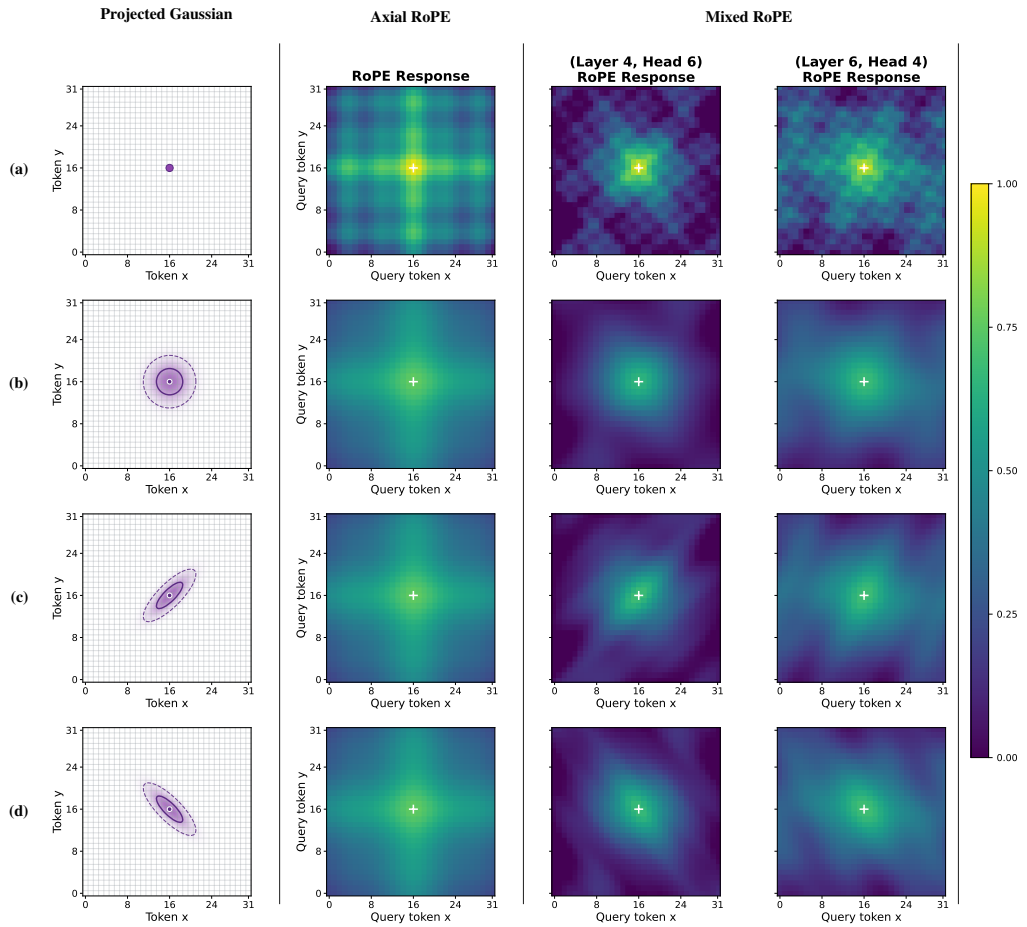


Figure 8: Projected Gaussian supports and their responses under axial and mixed rotary operators. We set $\mathbf{q} = \mathbf{k} = \mathbf{1}/\sqrt{d}$ to isolate the positional response. Panel (a) shows a point hypothesis; panels (b) to (d) have the same marginal variances Σ_{uu} and Σ_{vv} but different off-diagonal covariance Σ_{uv} and Σ_{vu}. Across these panels, the axial operator gives the same response, whereas the mixed operator responds to the covariance between image coordinates.

Both axial and mixed rotary operators (Heo et al., 2024) compute each rotation angle as $\boldsymbol{\omega}^\top \mathbf{u}$ and thus support the Gaussian integration derived in the main paper. They differ in the choice of $\boldsymbol{\omega}$ and the covariance terms used in attenuation. For $\boldsymbol{\omega} = (\omega_u, \omega_v)^\top$, the phase variance is given by

$$\boldsymbol{\omega}^\top \boldsymbol{\Sigma} \boldsymbol{\omega} = \omega_u^2 \Sigma_{uu} + 2\omega_u \omega_v \Sigma_{uv} + \omega_v^2 \Sigma_{vv}. \tag{14}$$

Axial rotary operators compute each rotation angle from either the horizontal or vertical coordinate, using $\boldsymbol{\omega} = (\omega_u, 0)^\top$ or $(0, \omega_v)^\top$. The phase variance is then given by

$$\boldsymbol{\omega}^\top \boldsymbol{\Sigma} \boldsymbol{\omega} = \omega_u^2 \Sigma_{uu} \quad \text{or} \quad \omega_v^2 \Sigma_{vv}, \tag{15}$$

so attenuation depends only on the corresponding marginal variance. Mixed rotary operators instead combine both coordinates, allowing their attenuation to use the cross-covariance term. They can therefore distinguish projected Gaussians with identical marginal variances but different orientations, as illustrated in Figure 8.

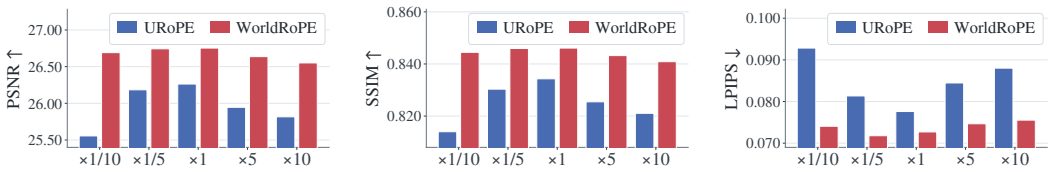
As shown in Table 8, WorldRoPE gains most consistently from mixed rotary operators, while RayRoPE and URoPE show limited or inconsistent improvements. This suggests that mixed rotary operators make fuller use of the cross-coordinate covariance in our projected Gaussians.

Table 8: Comparison of axial and mixed rotary operators for novel view synthesis on the RE10K and Objaverse datasets. Parenthesized values denote changes from the axial variant.

Method	**Operator**	**RE10K** (Zhou et al., 2018)			**Objaverse** (Deitke et al., 2023)		
		PSNR ↑	**SSIM** ↑	**LPIPS** ↓	**PSNR** ↑	**SSIM** ↑	**LPIPS** ↓
RayRoPE (Wu et al., 2026)	Axial	26.17	0.833	0.081	22.25	0.892	0.121
	Mixed	26.14 (−0.03)	0.832 (−0.001)	0.082 (+0.001)	22.33 (+0.08)	0.894 (+0.002)	0.118 (−0.003)
URoPE (Xie et al., 2026)	Axial	26.26	0.834	0.078	21.99	0.890	0.128
	Mixed	26.27 (+0.01)	0.834 (0.000)	0.079 (+0.001)	22.03 (+0.04)	0.890 (0.000)	0.126 (−0.002)
WorldRoPE (ours)	Axial	26.75	0.846	0.073	22.52	0.896	0.113
	Mixed	**26.84** (+0.09)	**0.848** (+0.002)	**0.072** (−0.001)	**22.71** (+0.19)	**0.898** (+0.002)	**0.108** (−0.005)

A.2 Sensitivity to depth initialization.

Figure 9 examines sensitivity to the depth configuration. We scale the predefined depth anchors of URoPE and the initial depth biases of WorldRoPE from 0.1 to 10 times their default values. WorldRoPE maintains higher PSNR and SSIM and lower LPIPS at every tested scale, with smaller performance variations. In particular, reducing the scale to 0.1 substantially degrades URoPE, while WorldRoPE remains close to its default performance. These results indicate that the learned depth hypotheses are less sensitive to their initial placement.



(a) PSNR ↑ (b) SSIM ↑ (c) LPIPS ↓

Figure 9: Sensitivity to initialization scale on RE10K. The scale multiplies the initial anchors of URoPE and the initial bias of WorldRoPE, with $\times 1$ denoting the default initialization. WorldRoPE remains comparatively stable across a broad range of initialization scales.

A.3 VISUALIZATION OF WORLD-SPACE GAUSSIAN PRIMITIVES

Figure 10 visualizes the predicted primitives and their projections into the query views. The projected supports occupy geometrically plausible regions, allowing WorldRoPE to produce stronger rotary responses for key and query tokens that observe the same 3D region across views.



Figure 10: Visualization of the probabilistic world-space primitive projection process. For each example, tokens from the key views are unprojected into probabilistic world-space primitives using the predicted depth and uncertainty. Each primitive uses the mean RGB of its source patch as color and its mixture weight as opacity for visualization. The primitives are then projected into the query views and compared with the corresponding ground-truth views.

A.4 LAYER-WISE GEOMETRY PREDICTIONS

Figure 11 visualizes the world-space primitives predicted across successive Transformer layers. Scaling σ_d by 0.2 yields a clearer spatial structure than using the predicted uncertainty scale directly. Although rotary integration and explicit visualization favor different uncertainty scales, a simple rescaling of σ_d produces a coherent spatial structure.

The predicted geometry becomes progressively more coherent through the intermediate layers, while the final layer shows a distinct pattern, possibly reflecting task-specific specialization near the output head. Unless otherwise specified, we use the Layer 4 primitives with $\sigma_d \times 0.2$ for visualization throughout the paper.

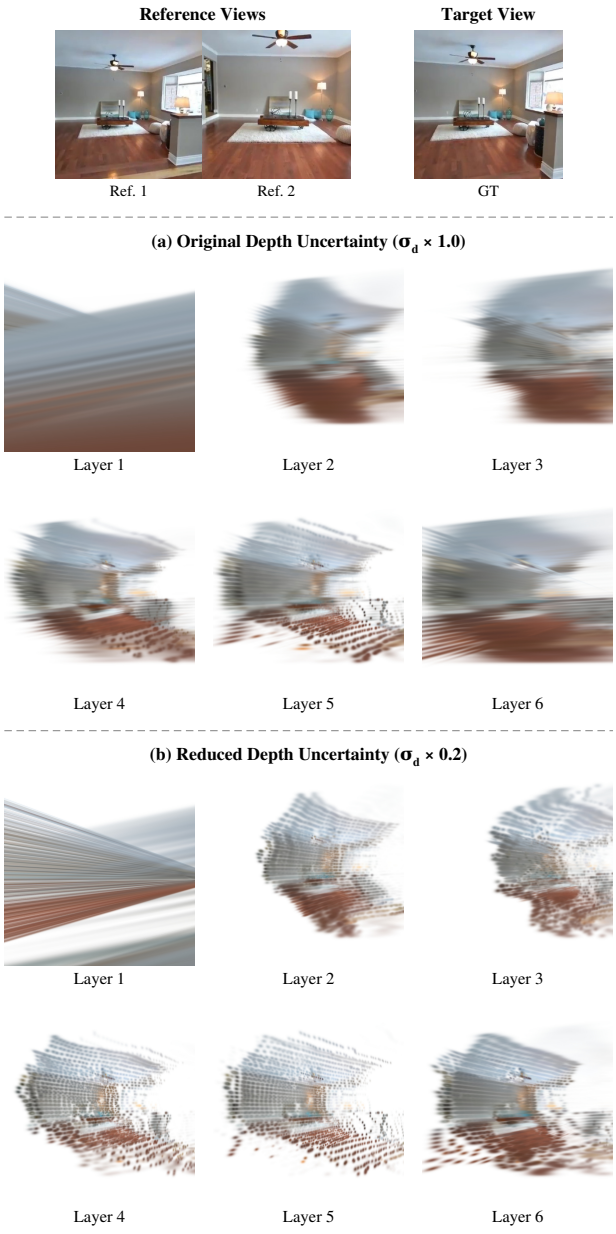


Figure 11: Layer-wise visualization of predicted world-space Gaussian primitives. We render the Gaussian primitives predicted at Layers 1–6 using the original depth uncertainty ($\sigma_d \times 1.0$) and a reduced uncertainty ($\sigma_d \times 0.2$). For visualization, each primitive is colored by the mean RGB of its corresponding image token, with its mixture weight used as opacity.

A.5 SENSITIVITY TO CAMERA VARIATIONS

Figure 12 evaluates robustness under varying multi-view context and camera conditions. In (a)–(c), although all models are trained with two reference views, WorldRoPE remains stable as the number of input views increases and achieves the strongest performance under multi-view inference. In contrast, PRoPE and RayRoPE degrade substantially with larger view counts. Their default implementations apply inter-camera distance normalization during training, which may contribute to this degradation, whereas URoPE and WorldRoPE do not use this normalization.

In (d)–(f), WorldRoPE maintains strong performance under image zoom-in, and similar robustness is observed under reference-camera rotation noise in (g)–(i). WorldRoPE also maintains the strongest performance across reference-camera distance ranges in (j)–(l). Overall, WorldRoPE remains robust across variations in multi-view context and camera geometry.

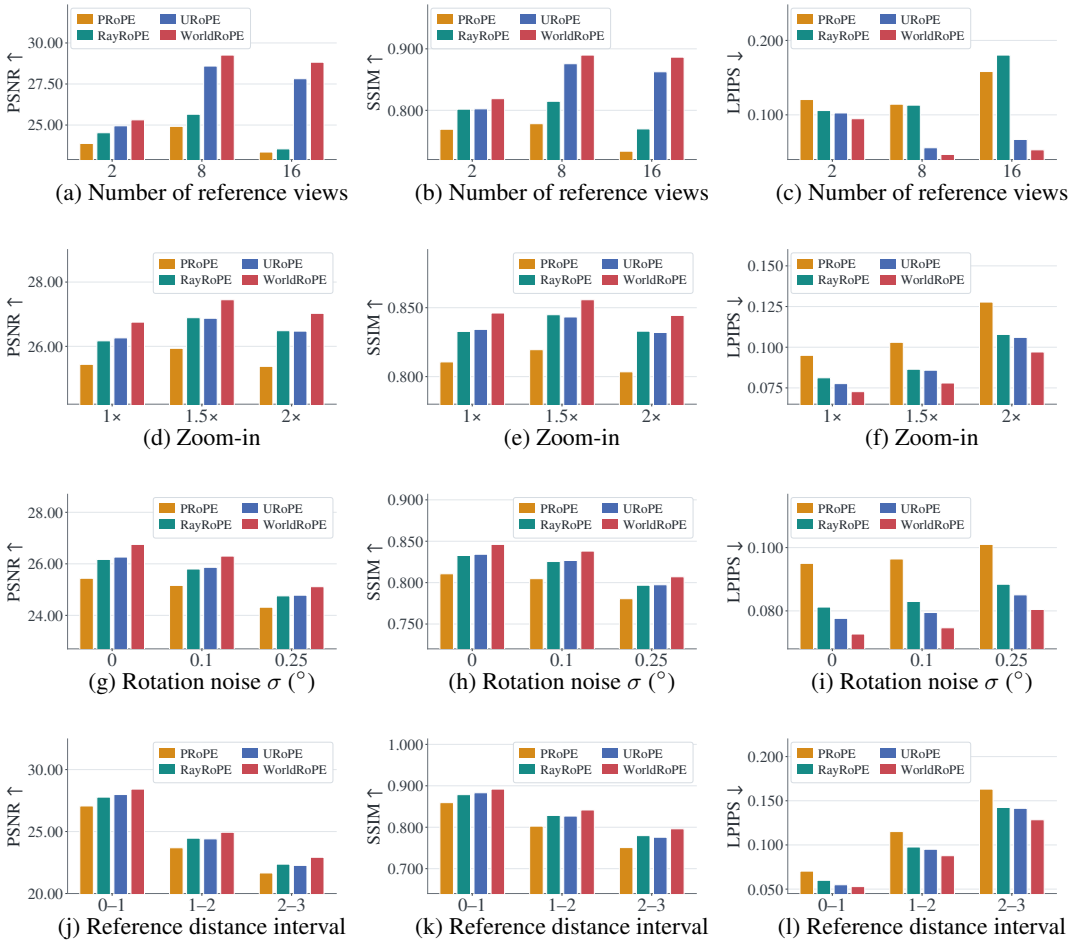


Figure 12: Robustness evaluation on RE10K under varying multi-view context and camera conditions. (a)–(c) vary the number of reference views at inference time. (d)–(f) evaluate image zoom-in. (g)–(i) evaluate reference-camera rotation noise. (j)–(l) evaluate reference-camera distance. Higher PSNR and SSIM and lower LPIPS indicate better performance.

A.6 Token-Adaptive Temperature

We examine whether the token-adaptive temperature in WorldRoPE can be replaced by a learned scalar without modeling positional covariance. Each key token i predicts depth hypotheses $\mu^m_{k,i}$ and mixture weights $\alpha^m_{k,i}$ as in WorldRoPE, but does not model depth uncertainty $\sigma^m_{k,i}$ or patch extent, which determine the temperature. Instead, it predicts a temperature $\tau^m_{k,i}$ for each primitive, $\{(\alpha^m_{k,i}, \mu^m_{k,i}, \tau^m_{k,i})\}_{m=1}^M = \Phi'(\mathbf{f}_{k,i})$. As a result, the learned temperature is shared across all rotary blocks, whereas WorldRoPE derives a separate temperature from the predicted covariance for each rotary frequency. Table 9 shows that the geometry-aware token-adaptive temperature in WorldRoPE provides a clear performance gain over a learned token-adaptive temperature.

Table 9: Ablation of token-adaptive temperature on RE10K.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
w/o Temperature	26.30	0.835	0.082
w/ Learned Temperature	26.33	0.836	0.082
WorldRoPE	**26.75**	**0.846**	**0.073**

A.7 Impact of Rotary Embedding on Value and Output

Following GTA (Miyato et al., 2024), we also apply positional transformations to values and outputs. Given the attention weights $A^{q\leftarrow k}_{ji}$ from Equation (10), let $\mathbf{v}_i$ denote the value of key token i and $\mathbf{o}_j$ the attention output for query token j. The value and output transformations are then given by

Table 10: Ablation of value and output rotations on RE10K.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
w/o V/O rotation	26.39	0.838	0.079
WorldRoPE	**26.75**	**0.846**	**0.073**

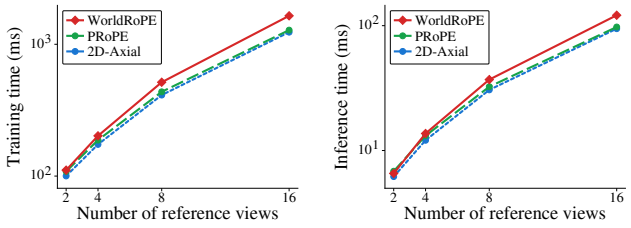
$$\mathbf{o}_j = \mathcal{R}(\mathbf{u}_{q,j})^\top \sum_{k,i} A^{q\leftarrow k}_{ji}\, \overline{\mathcal{R}}_{k,i\to q}\, \mathbf{v}_i, \tag{16}$$

where the factor $\mathcal{R}(\mathbf{u}_{q,j})^\top$ corresponds to the *Inverse Rot.* block in Figure 2. Table 10 quantifies the effect of the value and output rotations.

A.8 Computational Cost

WorldRoPE adds only 55K trainable parameters (+0.23%) to the 25M LVSM model. These additional parameters are introduced by layer-wise linear heads that predict depth means, depth standard deviations, and mixture weights.

We measure training and inference time on a single NVIDIA A6000 GPU. Training uses a batch size of 4, while inference is measured with one target view. All models are benchmarked with the full model compiled using `torch.compile`.



(a) Training time (b) Inference time

Figure 13: Training and inference time as the number of reference views increases. 2D-Axial denotes standard 2D image-space RoPE.

Figure 13 shows that the runtime of WorldRoPE grows faster than that of 2D-Axial and PRoPE as the number of reference views increases. To integrate with existing image-token attention as a plug-in, we project the predicted geometry into each query view at every layer, resulting in $\mathcal{O}(N^2)$ geometric computation for N views. The expected rotary operator also has a closed form for Gaussian positions in 3D. Future models that operate directly in shared 3D space could therefore use the same formulation without repeated view-pair projections.

B Derivations and Proofs

We follow the notation of the main paper. Consider a patch j in a query view q and a patch i in a key view k. Their coordinates are $\mathbf{u}_{q,j} = (u_{q,j}, v_{q,j})^\top$ and $\mathbf{u}_{k,i} = (u_{k,i}, v_{k,i})^\top$. For each view r, let $\mathbf{K}_r$ and $(\mathbf{R}_r, \mathbf{t}_r)$ denote its intrinsic and extrinsic camera parameters. The query coordinate $\mathbf{u}_{q,j}$ is deterministic, while $p_{k,i\to q}$ describes the projected positional distribution of key token i. We fix one rotary frequency $\boldsymbol{\omega}$ and omit its block index unless assembling the full operator. For generic single-primitive derivations, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ denote the projected center and covariance of a distribution in $\mathbb{R}^2$.

B.1 Expected Rotary Operators

Closed-form expectation. For a single projected Gaussian primitive, let $\mathbf{u} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in $\mathbb{R}^2$, suppressing token and view indices. Since each rotary block consists of sine and cosine terms, its expectation is

$$\mathbb{E}[R(\boldsymbol{\omega}^\top \mathbf{u})] = \begin{bmatrix} \mathbb{E}[\cos(\boldsymbol{\omega}^\top \mathbf{u})] & -\mathbb{E}[\sin(\boldsymbol{\omega}^\top \mathbf{u})] \\ \mathbb{E}[\sin(\boldsymbol{\omega}^\top \mathbf{u})] & \mathbb{E}[\cos(\boldsymbol{\omega}^\top \mathbf{u})] \end{bmatrix}. \tag{17}$$

We compute these terms together using Euler's identity, $e^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta$. The characteristic function of the Gaussian gives

$$\begin{aligned} \mathbb{E}[e^{\mathrm{i}\boldsymbol{\omega}^\top \mathbf{u}}] &= \int_{\mathbb{R}^2} e^{\mathrm{i}\boldsymbol{\omega}^\top \mathbf{u}} \mathcal{N}(\mathbf{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\mathbf{u} \\ &= \exp\!\left(-\tfrac{1}{2}\boldsymbol{\omega}^\top \boldsymbol{\Sigma} \boldsymbol{\omega}\right) e^{\mathrm{i}\boldsymbol{\omega}^\top \boldsymbol{\mu}}. \end{aligned} \tag{18}$$

Taking the real and imaginary parts yields

$$\begin{aligned} \mathbb{E}[\cos(\boldsymbol{\omega}^\top \mathbf{u})] &= a\cos(\boldsymbol{\omega}^\top \boldsymbol{\mu}), \\ \mathbb{E}[\sin(\boldsymbol{\omega}^\top \mathbf{u})] &= a\sin(\boldsymbol{\omega}^\top \boldsymbol{\mu}), \end{aligned} \qquad a := \exp\!\left(-\tfrac{1}{2}\boldsymbol{\omega}^\top \boldsymbol{\Sigma} \boldsymbol{\omega}\right). \tag{19}$$

Both terms share the same attenuation a. Substituting them into Equation (17) therefore gives

$$\mathbb{E}[R(\boldsymbol{\omega}^\top \mathbf{u})] = aR(\boldsymbol{\omega}^\top \boldsymbol{\mu}). \tag{20}$$

Thus, the expected block is a rotation at the projected center multiplied by covariance-dependent attenuation. By linearity of expectation, combining the primitive contributions with weights $\alpha_{k,i}^m$ gives Equation (9).

Equivalent single-rotation form of the mixture. Although the rotary operator combines contributions from M Gaussian components, their weighted sum can be represented exactly by a single scaled rotation for each rotary frequency block.

For a fixed rotary frequency $\boldsymbol{\omega}$, recall from Eq. (9) that $\overline{R}_{k,i\to q} = \sum_{m=1}^{M} \alpha_{k,i}^m a_{k,i\to q}^m R(\boldsymbol{\omega}^\top \boldsymbol{\mu}_{k,i\to q}^m)$, where $a_{k,i\to q}^m = \exp(-\frac{1}{2}\boldsymbol{\omega}^\top \boldsymbol{\Sigma}_{k,i\to q}^m \boldsymbol{\omega})$. Let $\theta_m = \boldsymbol{\omega}^\top \boldsymbol{\mu}_{k,i\to q}^m$. The combined cosine and sine coefficients are given by

$$C_{k,i\to q} = \sum_{m=1}^{M} \alpha_{k,i}^m a_{k,i\to q}^m \cos\theta_m, \qquad S_{k,i\to q} = \sum_{m=1}^{M} \alpha_{k,i}^m a_{k,i\to q}^m \sin\theta_m. \tag{21}$$

The mixture operator can then be rewritten as follows:

$$\overline{R}_{k,i\to q} = \begin{bmatrix} C_{k,i\to q} & -S_{k,i\to q} \\ S_{k,i\to q} & C_{k,i\to q} \end{bmatrix} = \bar{a}_{k,i\to q} R(\bar{\theta}_{k,i\to q}). \tag{22}$$

The effective attenuation and phase are given by

$$\bar{a}_{k,i\to q} = \sqrt{C_{k,i\to q}^2 + S_{k,i\to q}^2}, \qquad \bar{\theta}_{k,i\to q} = \operatorname{atan2}(S_{k,i\to q}, C_{k,i\to q}). \tag{23}$$

When $\bar{a}_{k,i\to q} = 0$, the operator is zero and the phase is arbitrary. Thus, for each rotary frequency block, the Gaussian mixture is exactly equivalent to a single scaled rotation with effective attenuation $\bar{a}_{k,i\to q}$ and phase $\bar{\theta}_{k,i\to q}$.

Extension to probabilistic queries and keys. WorldRoPE naturally generalizes to the case where both query and key positions are probabilistic. To describe their rotary interaction, consider independent positions $\mathbf{u} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{u}' \sim \mathcal{N}(\boldsymbol{\mu}', \boldsymbol{\Sigma}')$ in a common coordinate system. Their relative position is given by

$$\mathbf{u} - \mathbf{u}' \sim \mathcal{N}(\boldsymbol{\mu} - \boldsymbol{\mu}',\, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}')\,, \tag{24}$$

which allows us to compute its expected rotation using the same closed-form expression. Under the independence assumption, this expectation is equivalently obtained from the individual expected operators as follows:

$$\mathbb{E}[\mathcal{R}(\mathbf{u} - \mathbf{u}')] = \big(\mathbb{E}[\mathcal{R}(\mathbf{u}')]\big)^\top \mathbb{E}[\mathcal{R}(\mathbf{u})]. \tag{25}$$

Derivation of the sensitivity bound. For a projected Gaussian with attenuation a, consider perturbing its projected center $\boldsymbol{\mu}$ by $\boldsymbol{\delta}$. Then,

$$\begin{aligned}
&\big\| aR\big(\boldsymbol{\omega}^\top(\boldsymbol{\mu} + \boldsymbol{\delta})\big) - aR\big(\boldsymbol{\omega}^\top \boldsymbol{\mu}\big) \big\|_2 && (26)\\
&\quad = a \left\| R\big(\boldsymbol{\omega}^\top \boldsymbol{\mu}\big) \left[R\big(\boldsymbol{\omega}^\top \boldsymbol{\delta}\big) - \mathbf{I} \right] \right\|_2 && (27)\\
&\quad = a \left\| R\big(\boldsymbol{\omega}^\top \boldsymbol{\delta}\big) - \mathbf{I} \right\|_2 && (28)\\
&\quad = 2a \left| \sin\left(\frac{\boldsymbol{\omega}^\top \boldsymbol{\delta}}{2} \right) \right| \le a \left| \boldsymbol{\omega}^\top \boldsymbol{\delta} \right|. && (29)
\end{aligned}$$

B.2 Compatibility with Lie Group Theory

Lie-group-based rotary encodings (Ostmeier et al., 2025; Yu et al., 2025) parameterize rotations in high-dimensional feature spaces through matrix exponentials; we consider $\mathcal{R}(\mathbf{u}) = \exp(\mathbf{A}u + \mathbf{B}v)$, where $\mathbf{u} = (u, v)^\top$ and $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d \times d}$ are commuting skew-symmetric generators.

WorldRoPE can be applied to these rotations by integrating over the uncertainty in positional coordinates. For $\mathbf{u} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the expected rotation matrix is given by

$$\begin{aligned}
\mathbb{E}[\mathcal{R}(\mathbf{u})] &= \int_{\mathbb{R}^2} e^{\mathbf{A}u + \mathbf{B}v} \mathcal{N}(\mathbf{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\mathbf{u} && (30)\\
&= e^{\mathbf{A}\mu_u + \mathbf{B}\mu_v} \mathbb{E}\big[e^{\mathbf{A}\delta u + \mathbf{B}\delta v}\big] && (31)\\
&= e^{\mathbf{A}\mu_u + \mathbf{B}\mu_v} \exp\big(\tfrac{1}{2}\mathbb{E}\big[(\mathbf{A}\delta u + \mathbf{B}\delta v)^2\big]\big) && (32)\\
&= e^{\mathbf{A}\mu_u + \mathbf{B}\mu_v} \exp\Big(\tfrac{1}{2}\big[\Sigma_{uu}\mathbf{A}^2 + 2\Sigma_{uv}\mathbf{A}\mathbf{B} + \Sigma_{vv}\mathbf{B}^2\big]\Big). && (33)
\end{aligned}$$

As in the main paper, the mean determines the rotation, while the covariance determines the attenuation due to positional uncertainty. In particular, the term $2\Sigma_{uv}\mathbf{A}\mathbf{B}$ captures correlations between the two coordinate axes, as in mixed RoPE (Section A.1).

Comparison with uniform integration. For a one-dimensional interval as used in RayRoPE (Wu et al., 2026), consider $\mathbf{u} = \boldsymbol{\mu} + t\mathbf{s}$, where $t \sim \mathrm{Unif}[-1, 1]$ and $\mathbf{s} = (s_u, s_v)^\top$ denotes the interval half-extent. With $\mathbf{H} = \mathbf{A}s_u + \mathbf{B}s_v$, uniform integration gives

$$\mathbb{E}_{\mathrm{Unif}}[\mathcal{R}(\mathbf{u})] = e^{\mathbf{A}\mu_u + \mathbf{B}\mu_v} \operatorname{sinhc}(\mathbf{H}), \tag{34}$$

where $\operatorname{sinhc}(\mathbf{X}) = \sum_{n=0}^{\infty} \mathbf{X}^{2n}/(2n+1)!$. The uniform expectation depends on the interval direction and extent, whereas the Gaussian expectation in Equation (33) incorporates the positional covariance $\boldsymbol{\Sigma}$, including variation along both coordinate axes and their correlation. Within each planar rotation block, uniform integration yields a sinc factor that can change sign, while Gaussian integration yields positive exponential attenuation. Training also requires differentiating the expected rotary operator with respect to the learned generators. For invertible $\mathbf{H}$, the uniform factor can be written as $\mathbf{H}^{-1}\sinh(\mathbf{H})$. Direct differentiation of this form introduces inverse-dependent terms that can be numerically sensitive near singularity. In contrast, WorldRoPE uses the Gaussian expectation in Equation (33), avoiding this inverse-based differentiation and reducing potential numerical sensitivity near singular configurations.

B.3 Gaussian Primitive Construction

A finite image patch observes a spatial region whose extent cannot be represented by its central ray alone. Mip-NeRF addresses this geometric property by approximating conical frustums with Gaussians (Barron et al., 2021). We follow the same perspective of representing spatial extent with Gaussian moments, but construct our primitives from a rectangular patch footprint and predicted depth uncertainty. Below, we derive the mean and covariance in Equation (3) through first-order unprojection.

Patch and depth offsets. Let $\boldsymbol{\epsilon} \in \mathbb{R}^2$ denote the zero-mean pixel offset from the patch center, with covariance $\mathbf{C}$. For a uniformly weighted rectangular footprint of size $\Delta_u \times \Delta_v$,

$$\mathbf{C} = \mathrm{diag}\left(\frac{\Delta_u^2}{12}, \frac{\Delta_v^2}{12}\right). \tag{35}$$

For depth hypothesis m, write $D_{k,i}^m = \mu_{k,i}^m + \delta d$, where $\mathbb{E}[\delta d] = 0$ and $\mathbb{E}[(\delta d)^2] = (\sigma_{k,i}^m)^2$. Assuming independent pixel and depth offsets, the camera-space position is

$$\mathbf{X}_{k,i}^m = (\mu_{k,i}^m + \delta d)(\mathbf{r}_{k,i} + \mathbf{B}_k \boldsymbol{\epsilon}), \tag{36}$$

where $\mathbf{r}_{k,i} = \mathbf{K}_k^{-1}(u_{k,i}, v_{k,i}, 1)^\top$ and $\mathbf{B}_k$ contains the first two columns of $\mathbf{K}_k^{-1}$.

First-order Gaussian approximation. Linearizing the unprojection omits the product $\delta d\, \mathbf{B}_k \boldsymbol{\epsilon}$, yielding

$$\mathbf{X}_{k,i,\mathrm{lin}}^m = \mu_{k,i}^m \mathbf{r}_{k,i} + \mu_{k,i}^m \mathbf{B}_k \boldsymbol{\epsilon} + \mathbf{r}_{k,i} \delta d. \tag{37}$$

We define the Gaussian primitive by matching the mean and covariance of this linearized position. Since both offsets have zero mean, its center is

$$\boldsymbol{\mu}_{k,i}^m := \mathbb{E}[\mathbf{X}_{k,i,\mathrm{lin}}^m] = \mu_{k,i}^m \mathbf{r}_{k,i}. \tag{38}$$

Subtracting this mean and expanding the outer product gives

$$\begin{aligned} \boldsymbol{\Sigma}_{k,i}^m &:= \mathrm{Cov}(\mathbf{X}_{k,i,\mathrm{lin}}^m) \\ &= \mathbb{E}\big[(\mu_{k,i}^m \mathbf{B}_k \boldsymbol{\epsilon} + \mathbf{r}_{k,i}\delta d)(\mu_{k,i}^m \mathbf{B}_k \boldsymbol{\epsilon} + \mathbf{r}_{k,i}\delta d)^\top\big] \\ &= (\mu_{k,i}^m)^2 \mathbf{B}_k \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] \mathbf{B}_k^\top + \mathbb{E}[(\delta d)^2] \mathbf{r}_{k,i}\mathbf{r}_{k,i}^\top \\ &\quad + \mu_{k,i}^m \mathbf{B}_k \mathbb{E}[\boldsymbol{\epsilon}\, \delta d] \mathbf{r}_{k,i}^\top + \mu_{k,i}^m \mathbf{r}_{k,i} \mathbb{E}[\delta d\, \boldsymbol{\epsilon}^\top] \mathbf{B}_k^\top. \end{aligned} \tag{39}$$

The offsets are independent, so $\mathbb{E}[\boldsymbol{\epsilon}\, \delta d] = \mathbb{E}[\boldsymbol{\epsilon}]\mathbb{E}[\delta d] = \mathbf{0}$. The cross terms therefore vanish, and substituting the offset moments yields

$$\boldsymbol{\Sigma}_{k,i}^m = (\mu_{k,i}^m)^2 \mathbf{B}_k \mathbf{C} \mathbf{B}_k^\top + (\sigma_{k,i}^m)^2 \mathbf{r}_{k,i}\mathbf{r}_{k,i}^\top. \tag{40}$$

This recovers Equation (3), with the two covariance terms capturing the patch footprint and depth uncertainty, respectively.

C Implementation Details

C.1 Common Implementation

Across all experiments, we implement each predictor Φ as a layer-wise linear head. Mixture weights are normalized using a softmax, depth standard deviations are parameterized exponentially, and depth means are clamped to be nonnegative. These predicted mixture weights, depth means, and depth standard deviations receive no direct auxiliary supervision and are learned end-to-end through the task-specific objective. Table 11 summarizes the implementation settings for each task.

C.2 Novel View Synthesis: LVSM (Jin et al., 2025)

Our experiments use the 25M-parameter LVSM architecture implemented in the URoPE codebase (Xie et al., 2026) [2]. The model consists of six Transformer layers with a hidden dimension of 768, 16 attention heads, and a feedforward dimension of 1024. Images have a resolution of 256×256 and are divided into 8×8 patches. Each training sample consists of two reference views and one supervised target view. We train the model for 80K steps on two GPUs with a global batch size of 16, using MSE loss and VGG perceptual loss with weights of 1 and 0.5, respectively.

For the positional-encoding baselines, we use the Plücker Ray, GTA, and PRoPE implementations provided in the URoPE codebase, while URoPE is evaluated using its native implementation. We implement RayRoPE following its official codebase. All methods use the same backbone and training protocol.

We train separate models on RealEstate10K (RE10K) (Zhou et al., 2018) and Objaverse (Deitke et al., 2023). For RE10K, we follow the train/test scene split used in the official RayRoPE codebase (Wu et al., 2026) [3]. For Objaverse, we use object renderings from eight camera directions spanning the full azimuth range.

For the mixed RoPE (Heo et al., 2024) variants, we initialize learnable frequency vectors with the same magnitudes as the corresponding original frequencies but with randomly sampled unit directions. For RayRoPE, Table 8 reports results from mixing the u and v coordinates without mixing inverse depth, since additionally mixing inverse depth reduced PSNR by 0.28 dB on RE10K and 1.07 dB on Objaverse.

C.3 Depth Estimation: UniMatch (Xu et al., 2023)

We use the official UniMatch codebase (Xu et al., 2023) [4]. Models are trained from scratch on the Scenes11 (Ummenhofer et al., 2017), SUN3D (Xiao et al., 2013), and RGBD (Sturm et al., 2012) datasets for 50K steps on two GPUs with a global batch size of 16. The Transformer has six layers, 144 feature channels, and three attention heads. Training uses 224×288 images, with evaluation at 240×320. Self-attention retains the 2×2 shifted-window structure, while cross-attention is global. Other settings follow the UniMatch codebase.

The geometry-aware encodings of GTA, PRoPE, URoPE, RayRoPE, and WorldRoPE in Table 6 are applied only to cross-attention, whereas the Plücker Ray baseline adds linearly projected Plücker coordinates to the image features. GTA, PRoPE, URoPE, and RayRoPE follow their official implementations, adapted to UniMatch.

For evaluation, we follow the official UniMatch protocol and use the same benchmark evaluation sets and preprocessing provided in its codebase. For unseen-data evaluation on ScanNet++ (Yeshwanth et al., 2023), we adopt the validation and test splits defined by MapAnything (Keetha et al., 2026). To account for scale differences between the training data and unseen ScanNet++, we align each prediction to the ground-truth scale using least squares before computing Abs Rel and RMSE.

[2] https://github.com/Applied-Intuition-Open-Source/URoPE

[3] https://github.com/Lucas-707/RayRoPE

[4] https://github.com/autonomousvision/unimatch

C.4 3D GEOMETRY ESTIMATION: VGGT (WANG ET AL., 2025A)

We initialize from the pretrained VGGT-1B model, which contains 24 alternating frame-wise and global self-attention stages with 1,024-dimensional features and 16 attention heads. Following the geometry-conditioning design of Pi3X (Wang et al., 2025b), we insert eight geometry-attention blocks with the same hidden dimension as the VGGT backbone, placing one after every third global-attention block. We fully train the eight added blocks and the camera and depth prediction heads, while applying rank-8 LoRA (Hu et al., 2022) to the corresponding eight pretrained global-attention blocks and keeping the remainder of the pretrained model frozen. We train on the ScanNet++ training split provided by MapAnything (Keetha et al., 2026), comprising 896 scenes. We evaluate on the corresponding 30-scene ScanNet++ test split.

Table 11: Summary of the task-specific implementations of WorldRoPE.

Setting	**Novel-view synthesis**	**Depth estimation**	**3D geometry estimation**
Backbone	LVSM-25M	UniMatch	VGGT-1B
Training data	RE10K or Objaverse	Scenes11, SUN3D, RGB-D SLAM	ScanNet++ (896 scenes)
Evaluation data	RE10K or Objaverse test split	Scenes11, SUN3D, RGB-D SLAM, and ScanNet++	ScanNet++ (30 held-out scenes)
Training views	2 references + 1 target	2 views	4 views
Resolution (train)	256×256	224×288	518×294
Resolution (eval.)	256×256	240×320	518×294
Components (M)	4	3	4
Training objective	$\mathcal{L}_{\text{MSE}} + 0.5\mathcal{L}_{\text{VGG}}$	$20\mathcal{L}_{\text{inv}} + 20\mathcal{L}_{\nabla\text{inv}}$	Confidence-weighted robust log-depth and world-point regression
Training steps	80K	50K	50K
GPUs	2	2	1
Batch size	16	16	2
Training precision	FP16	BF16	BF16

D Additional Qualitative Results

D.1 Novel view synthesis

Figure 14 presents additional novel-view synthesis results on RE10K (Zhou et al., 2018) and Objaverse (Deitke et al., 2023).



Figure 14: Qualitative comparison of novel-view synthesis on RE10K and Objaverse. The top five rows show RE10K examples, and the bottom five rows show Objaverse examples. The first two columns contain the reference views, followed by the target ground truth and predictions from the compared methods.

Figure 15 presents qualitative novel-view synthesis results for the axial and mixed RoPE (Heo et al., 2024) variants on RE10K (Zhou et al., 2018) and Objaverse (Deitke et al., 2023).



Figure 15: Qualitative comparison of the axial and mixed RoPE variants on RE10K and Objaverse. The top five rows show RE10K examples, and the bottom five rows show Objaverse examples. For each positional encoding, the axial and mixed RoPE variants are shown side by side.

D.2 DEPTH ESTIMATION

Figure 16 presents qualitative depth estimation results on Scenes11 (Ummenhofer et al., 2017), SUN3D (Xiao et al., 2013), RGBD (Sturm et al., 2012), and unseen ScanNet++ (Yeshwanth et al., 2023).

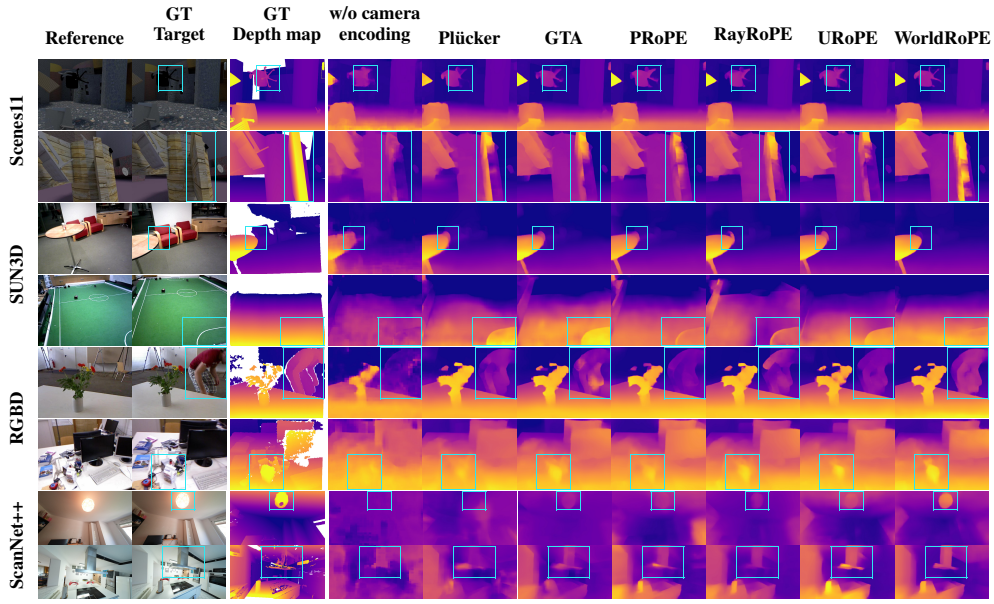


Figure 16: Qualitative comparison of depth estimation on Scenes11, SUN3D, RGBD, and unseen ScanNet++. The row groups follow this dataset order from top to bottom. The first three columns contain the reference image, target image, and ground-truth depth, followed by predictions from the compared methods.